Methods in Computational Linguistics I
LING 78100
Fall 2010

Homework #1 - Introduction to Python and NLTK
Due at 11:50 on September 17

1. NLTK Questions

A) [5] How many words are there in text5?

B) [5] How many times does 'heaven' appear in 'Moby Dick'? How many times in 'Genesis'?

C) [10] We have seen how to represent a sentence as a list of words, where each word is a sequence of characters. What does `sent1[2][2]` do? Why? Experiment with other index values.

D) [10] Consider the following Python expression: `len(set(text4))`. State the purpose of this expression. Describe the two steps involved in performing this computation.

E) [10] Lexical Diversity - Calculate the lexical diversity of the 8 texts. Recall: lexical diversity is defined as the ratio of the size of the text to the size of the vocabulary

F) Define sent to be the list of words `['she', 'sells', 'sea', 'shells', 'by', 'the', 'sea', 'shore']`. Now write code to perform the following tasks:

1. [5] Print all words beginning with *sh*
2. [5] Print all words longer than four characters

G) [10] What does the following Python code do? `sum([len(w) for w in text1])` Can you use it to work out the average word length of a text?

2. Open-ended Question

Play with the `text.generate()` command.

A) [10] Generate some random text based on 3 NLTK texts including those that you identified to have the largest and smallest lexical diversity. Are any of these believable sentences from the original text? Does the length of the generated text make a difference?
B) [5] Is there a difference between the quality of the texts based on lexical diversity? (Note, it's ok if there isn't.)
C) [5] How do you think these random texts are constructed?

D) [5] What are the biggest give-aways that these texts are not part of the original texts.
E) [15] How would you improve the construction of random texts that are similar to a given text?